My primary research interest lies in developing reliable language and foundation models. Recent language models (LMs) have acquired vast amounts of knowledge. However, they remain black boxes with limited interpretability and structure, making it difficult to systematically understand and expand the boundaries of their abilities. LMs also lack robustness to noisy data and overly rely on surface features, leading to unreliable outputs. These issues motivate my two research questions:

1. How can we facilitate reliable generalization with better **understanding** and **modularity**? [1]
2. How can we enhance LMs' factuality by inducing more **robust representation learning**? [2, 3, 4]

§1 Facilitate Reliable Generalization with Better Understanding and Modularity

As LMs grow in complexity, their lack of transparency and structure makes it challenging to understand their behaviors and ensure reliable generalization. I believe it is essential to build a deeper understanding of models' behaviors and to develop **modular** systems with clearer component specializations.

To build this understanding, key questions arise. **How do current architecture and training methods shape LMs' internal mechanisms and generalizability for different tasks?** Are the mechanisms formed during pretraining or post-training, and does post-training enhance expressiveness or constrain generalization? Can models effectively share mechanisms across tasks to achieve compositional generalization? Answering these questions can help identify LMs' capabilities and limitations and lay a foundation for developing better learning paradigms.

In our upcoming ACL'25 paper, we investigated how LMs use components like attention and MLPs to handle multi-answer factual queries [1]. By decoding and intervening on attention outputs, we found that attention uses subject tokens to promote the subject and previous answer tokens to suppress repetition. Meanwhile, MLPs use subject information from attention to retrieve answers and amplify suppression signals. It is fascinating to see components coordinate with each other to answer the queries. However, I found the coordination imperfect: components can *negatively interfere* with others, such as wrongly promoting one answer while others suppress it. This issue can undermine effective coordination among components and reliable generalizations.

Motivated by this observation, I want to make components more modular and specialized. Seeing components already attend to specific tokens to achieve different goals, I want to design architectures and objectives that **route information to specialized components for specific goals**, such as extracting semantics from varied tokens of the same concept (§2.1), encoding and retrieving domain knowledge with structures (§2.2), and removing undesirable tokens like repetitions or harmful suggestions if needed. For instance, we could regularize components to focus on the most important tokens based on activation scores or predict token importance in advance and route them accordingly. By exploring this direction, I aim to encourage component specialization and reliable generalization.

§2 Enhance Factuality with Robust Representation Learning

LMs trained on noisy and unbalanced data can lack robustness and reliability. First, bound by tokenizers, LMs must learn separate features for semantically related tokens, adding **unnecessary noise** that increases LMs' sensitivity to input perturbations. Second, LMs trained on unbalanced data can **overly rely on surface features** and generate incorrect outputs. I aim to develop more robust, structured representation learning to tackle both issues.

§2.1 Reduce Noise in Representation Learning

LMs trained on web-scraped data can capture noisy features, making them prone to inaccurate responses even if the models encode factual knowledge [5, 9]. For instance, slight changes in prompt formatting can cause LMs to provide incorrect answers [11]. I approached this problem with methods external and internal to the models. Externally, we addressed this in our EMNLP'24 paper by combining the logits of a paraphraser and the target model to rewrite prompts to have lower perplexity, making them more familiar to the target model—something unattainable by having the model rephrase the prompts itself [3]. Internally, I improved LMs' robustness by maximizing the similarity among hidden states of paraphrased inputs in our ACL'24 paper [2]. It was beautiful to see the hidden states of inputs with various types of unseen perturbations were more aligned together.

Nevertheless, one observation still bothers me: LMs trained with standard next-token loss clustered hidden states of the same input with different perturbations into *distinct groups*, illustrating LMs' sensitivity to surface variations. This issue may stem from tokenization constraints, where variations like "Dog" and "Dogs" or minor spacing like "make" vs. "␣make" result in distinct embeddings in LLaMA and Mistral. While the flexibility to handle rare or unknown tokens is crucial, it also introduces fragmented and suboptimal representations that *distract LMs from capturing the structure and meaning of the inputs, waste parameters and data, and lead to undesirable outputs* [10]. I believe we need **more robust and efficient encoding of inputs to avoid wasting data and hidden subspaces for superficially orthogonal features**. I want to explore better tokenization schemes, compress intermediate representations, or create tokenizer-free LMs to improve models' robustness to rare or unseen tokens across domains and languages.

§2.2 Regularize and Structure Latent Representations to Mitigate Overreliance on Surface Features

LMs can overly rely on surface features to structure their latent spaces and generate errors, especially for long-tail data. I observed that even models like GPT4 can give misinformation based on the popularity or co-occurrence of input entities. In our ACL'23 paper, we mitigated surface biases in LMs by combining output logits and low-layer attention scores of a smaller auxiliary and our larger target model. Our method encourages the smaller model to capture malicious shortcuts while guiding the target model to fit more robust attention patterns. It was remarkable to see that our model learned to assign higher attention scores to more important tokens and form more robust representations.

However, this method may not precisely remove all spurious features across domains. Without a robust and structured latent space, models may suffer from *representation collapse and hallucination on long-tail data*, as well as *rippling errors when learning new data* [6, 7, 12]. To address these issues, I want to explore **regularization for structured latent representations** to enhance models' reliability and adaptability. One way to achieve this structure is to explicitly align LMs' internal representations with knowledge graph embeddings. However, I found directly imposing external structures on pretrained models challenging in my experiments. Given that LMs may already form locally linear representations for some structures of real-world data [14, 15], can we leverage existing structures or design objectives to induce a more structured hidden space to model human knowledge structures with fewer errors? Exploring these questions could help models form more accurate and updatable representations.

§3 PhD at XXX and Future Plans

[Why school & professors]

In the long term, I aspire to become a professor and lead a research lab. As a teaching assistant for machine learning and algorithms courses at USC, I found it deeply fulfilling to tailor my mentorship to support students with unique backgrounds and personalities. Inspired by my advisors and mentors, I aim to lead research that encourages deeper reflection on matters we have taken for granted.

References

[1] **Tianyi Lorena Yan**, and Robin Jia. "LLMs Combine Knowledge Recall and Copy Suppression to Answer One-to-Many Factual Queries. *To be submitted to ACL 2025.*

[2] **Tianyi Lorena Yan**, Fei Wang, James Y. Huang, Wenxuan Zhou, Fan Yin, Aram Galstyan, Wenpeng Yin, and Muhao Chen. Contrastive Instruction Tuning. *ACL 2024.*

[3] Qin Liu, Fei Wang, Nan Xu, **Tianyi Lorena Yan**, Tao Meng, and Muhao Chen. Monotonic paraphrasing improves generalization of language model prompting. *EMNLP 2024.*

[4] Fei Wang, James Y. Huang, **Tianyi Lorena Yan**, Wenxuan Zhou, and Muhao Chen. Robust natural language understanding with residual attention debiasing. *ACL 2023.*

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *ICLR 2023.*

[6] Jiachen Zhu, Katrina Evtimova, Yubei Chen, Ravid Shwartz-Ziv, and Yann LeCun. Variance-covariance regularization improves representation learning. arXiv preprint arXiv:2306.13292 (2023).

[7] Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *TACL 2024.*

[8] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language?. *ACL 2019.*

[9] Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On Large Language Models' Hallucination with Regard to Known Facts. *NAACL 2024.*

[10] Sander Land, and Max Bartolo. "Fishing for Magikarp: Automatically Detecting Under-trained Tokens in Large Language Models." *EMNLP 2024*.

[11] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *ICLR 2024.*

[12] Ryosuke Takahashi, Go Kamoda, Benjamin Heinzerling, Keisuke Sakaguchi, and Kentaro Inui. The Curse of Popularity: Popular Entities have Catastrophic Side Effects when Deleting Knowledge from Language Models. *NAACL 2024*.

[13] Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. Learning from others' mistakes: Avoiding dataset biases without modeling them. *ICLR 2021.*

[14] Templeton, et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet", Transformer Circuits Thread, 2024.

[15] Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. *ICLR 2024.*